Problem Set 3

SOC-GA 2332 Intro to Stats (Fall 2024)

Due: Friday, Nov. 22th, 11:59 pm

Contents

Instructions

1. Submit two files for each problem set. The first is a **R Markdown** (`.Rmd`) file that can be run without error from start to end. The second is a **PDF** rendered from your R Markdown file or created using LaTeX.
2. Name your files following this convention: `[Last Name]_ps3.Rmd` and `[Last Name]_ps3.pdf` (for example, `Jiang_ps3.Rmd`).
3. Both files should be submitted to the TA via e-mail (wj2068@nyu.edu) before the time specified above. Please email the TA at least three days before the due date if you need extensions with justified reasons. Please plan ahead and start early.
4. You are encouraged to discuss the problems with your classmates. But **the R Markdown and PDF files that you submit have to be created on your own**. Please do not ask for solutions from students in earlier cohorts.
5. Comment on your code wherever possible and explain your ideas in detail. You will get credits for showing the steps you take and for explaining your reasoning, even if you do not get the correct final result.

Prerequisite

Load multiple packages to your environment.

```
knitr::opts_chunk$set(echo = TRUE)

## load packages here
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Part 1 Multivariate Regression

`gss.csv` contains a dataset of the following variables:

Variable Name	Variable Detail
`id`	A respondent's unique ID
`abortion`	A respondent's answer to the question "Should a Woman be Able to Have Legal Abortion"; -1 = No, 0 = It depends, and 1 = Yes
`pid`	A respondent's party identification; 0 = Strong Democrat, 1 = Democrat, 2 = Democrat Leaning, 3 = Independent, 4 = Republican Leaning, 5 = Republican, 6 = Strong Republican
`linc`	A respondent's income (logged)
`female`	A dummy variable of sex; 1 = Female, 0 = Male
`college`	A dummy variable of college education; 1 = Yes, 0 = No
`mar`	A dummy variable of marital status; 1 = Married, 0 = Not married
`trump`	A dummy variable of voting for Trump in 2016; 1 = Yes, 0 = No

1. [5pts]Plot the distribution of `abortion` using a bar plot
2. [5pts]Using `abortion` as the outcome variable, fit Model 1 with the following predictors: `female`, `pid`, `college`, `linc`, `mar`; and Model 2 by adding an interaction between `female` and `pid` on Model 1.

- *Hint: treat both `abortion` and `pid` as continuous variables and fit regular OLS*

3. [5pts]Calculate heteroskedasticity-robust standard errors for Model 2; call this model Model 3.
4. [5pts]Create a regression table with three columns: Model 1, Model 2, and Model 3.
5. [10pts]Plot the predicted value of `abortion` of male and female respondents with different `pid` values. Be clear about which variables you fix at what values when generating the plot.
6. [5pts]Using both the tables and the plot, interpret the results.

Part 2 Logistic Regression

In Lab 8, we fitted a logistic regression model to predict the probability of one voting for Trump in the 2016 election. We also discussed potential problems of treating `pid` as a continuous numeric variable. Using the same data (`gss.csv`):

1. [5pts]Replicate the linear probability model and logistic model fitted in Lab 8, then fit a new logistic regression model using the same set of predictors but treat `pid` as an ordered categorical variable.

- *Hint: by ordered categorical variable, I mean you need to factorize `pid`, or create dummy categories*

2. [5pts]Create a regression table with three columns: Model 1 (linear probability model), Model 2 (logistic regression with `pid` as a numeric variable), and Model 3 (logistic regression with `pid` as an ordered

categorical variable).

3. [10pts]Plot the predicted probability of voting for Trump by `pid` based on Model 3 results (with 95% confidence intervals). Be clear about which variables you fix at what values when generating the plot. For your convenience, you should use the same values I used in Lab 8, *i.e.*, a male without college education with median income.

4. [5pts]Using the regression results above, interpret the coefficient(s) of `pid` across three models. Then, discuss how treating `pid` as categorical may (or may not) affect your findings.

Part 3 Causality

A study on COVID-19 constructed a "COVID risk factor" score based on the COVID infection rate of a given area (defined by zip code).

A researcher wants to estimate the effect of having a vaccination center in the area on that area's COVID risk factor score. She compiled a dataset that contains each area's COVID risk factor score and whether the area has a vaccination center. She then estimated the effect of having a vaccination center using the "naive estimator" we discussed in class.

You noted that the quality of information residents have about COVID and the vaccine can be a confounding variable that affects both the area's infection rate and whether there is a vaccination center in the area.

Assume that you are able to estimate the relationships this "informedness" confounder (`info`) and the original "vaccination center" predictor (`vaccine`) have with the COVID risk factor score (`covid_risk`), which can be simulated using the following code (`n` is sample size):

```
library("scales")
set.seed(1234) # set the same seed to ensure identical results
e = rnorm(n, 0, 0.5)
covid_risk = rescale( 0 - 7*vaccine - 2*info + e, to = c(0, 100))
```

1. [5pts]Import the data `covid.csv`. According to the potential outcome framework, create two new variables `y_c` and `y_t` where `y_c` is the value of the potential outcome of "risk factor" when the individual is not treated, and `y_t` is the value of the potential outcome of "risk factor" when the individual is treated. Explain your steps.

2. [5pts]Fill out the table below (round to 1 decimal points):

Group	Y^T	Y^C
Treatment Group ($D = 1$)	$E[Y^T \mid D = 1]$ =?	$E[Y^C \mid D = 1]$ =?
Control Group ($D = 0$)	$E[Y^T \mid D = 0]$ =?	$E[Y^C \mid D = 0]$ =?

3. [15pts]Estimate/calculate the following (write down the formula/equation you use):

i. The Naive Estimator of ATE
ii. The "true" Average Treatment Effect
iii. The "true" Treatment Effect on the Treated
iv. The "true" Treatment Effect on the Control
v. Selection Bias

4. [10pts]Write a non-technical, one-paragraph summary reporting your results in response to the above mentioned researcher who used the naive estimation. Imagine that you are explaining this to an audience who may not be familiar with the specific terminologies of the counterfactual framework (such as ATE or Treatment Effect on the Treated), but is interested in your substantive findings.